From Tombstones to Corpora: TSML for Research on Language, Culture, Identity and Gender Differences*

Oliver Streiter[a], Leonhard Voltmer[b], Yoann Goudin[c]

[a]National University of Kaohsiung, Taiwan, ostreiter@nuk.edu.tw
[b]European Academy Bolzano/Bozen, Italy, lvoltmer@eurac.edu
[c]Ecole des Hautes Etudes en Sciences Sociales, France, goudin@yahoo.com.fr

Abstract. Tombstone inscriptions represent a linguistic genre which yields insights in culture and language. Creating corpora from tombstones is thus a complementary approach for the study of languages and cultures. For the annotation of tombstone corpora, we propose TSML, the Tombstone-Markup-Language, developed during the massive annotation of Taiwanese tombstones and a number of tombstones from China, Indonesia and Europe. We discuss our conceptual framework in the annotation of tombstones and derive successively and present preliminary research data to show how the usefulness of the annotations. Finally, we will encourage researchers to participate in the specification of TSML to obtain soon an annotation language for annotations across cultures and languages.

Keywords: Tombstones, corpora, XML, TSML, Tombstone-Markup-Language, Taiwan.

1. From Tombstones to Corpora

Tombstone inscriptions represent a linguistic genre which, due to the moment it represents, allows for profound insights in culture and language. When the trivia of life don't matter anymore and the cullets of life are swept together in a few strokes in marble, language is frequently the only agent and the only trace in a battle between conflicting identities, social relations, conceptual systems, mythologies and religions.

Plate 1 to 4: From left to right, a sinicized aborigine tomb, a Japanese-style Han tomb, a christianized Han tomb and a de-sinicized aborigine tomb.

Tombstones can be found worldwide. Their form and content follow ethnic and religious traditions, cf. Rath 1986, Frembgen 1998,The Hindu 2005). Notwithstanding global trends, tombstone preserve very local customs. Even on a small island like Taiwan, tombstones in the North and South, East and West are different, blending the flavors of ethnic or religious traditions with local craftsmanship (cf. Clark 1989/1992). In addition, traditions in contact borrow from each other and create particular forms as shown in Plates 1 to 4.

Research on tombstones thus backs the study of language and culture. Creating a corpus of tombstones, as opposed to other research designs used with to tombstones, requires most

investments, but is also the most rewarding strategy. First, a corpus can reveal facts, such as local, temporal, ethnic, religious, social or gender-related differences, that cannot be learned from individual tombstones. In addition, a corpus, when properly balanced, paves the way to innumerable investigations beyond the initial motivation for the construction of the corpus. Third, a corpus with digital recordings such as photos as integral part can be continuously annotated, opening new perspectives with each new annotation. Finally, corpora from different resources can be used in comparative studies.

Tombstone corpora cannot bridge the gap between the rapid extinction of cultures and languages (Wurm 1991) and the missing research activities in language documentation. Although one might hope that tombstones will still be recoverable after the death of a culture, factors like urbanization, industrialization, tourism and construction work or acid rain threaten the existence of these mute witnesses. Although after 100 years one might still find individual tombstones, any systematic comparison across regions, ethnicities or time periods would be difficult. Our study on Taiwanese tombstones confirms the precarious state, showing a massive loss of tombstones of the time before 1950.

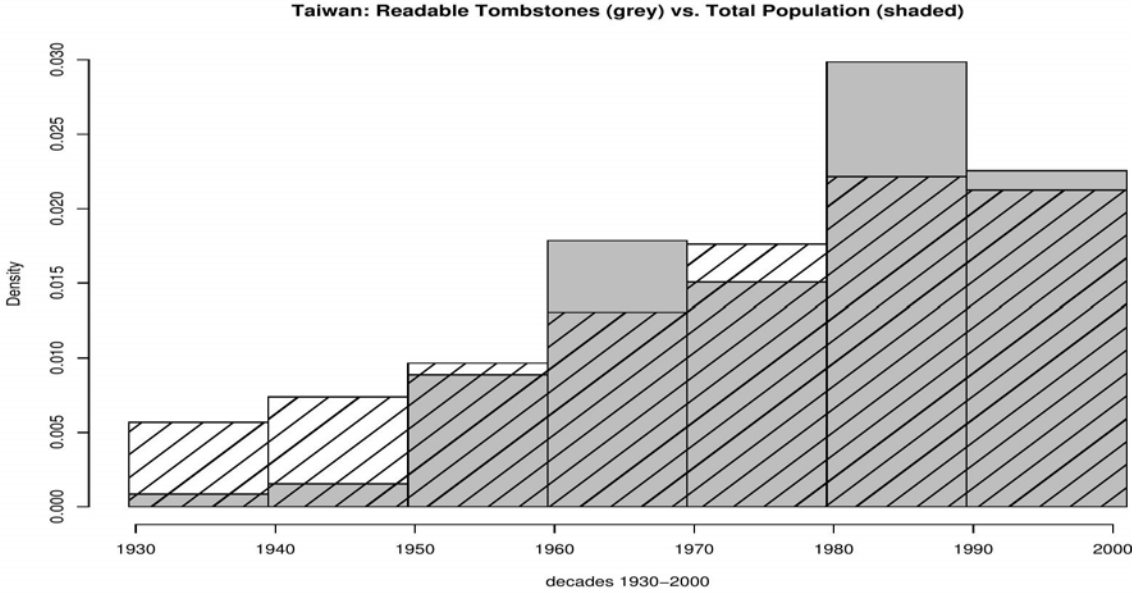

Figure 1: The density of tombstones compared to the density of the population in Taiwan through time. The lack of older tombstones cannot be explained through a smaller population in earlier times. Data are based on 3000 tombstones from 30 graveyards.

2. Annotating Tombstones

Different grass-root activities have sprung up, e.g. in the US and Australia, to preserve the cultural heritage of tombstones by photographing or transcribing them (e.g. http://www.rootsweb.com/~cemetery/). However, the nature of the transcriptions determines the use one can make of them, cf. Debartolo Carmack 2002. Unstructured transcriptions, for example, leave too much ambiguity for automatic analysis. The word 'Brown' might be a name or a color, 'Miller' a name or profession. 'Brown' thus should be annotated as 'name' and 'Miller' as 'profession'. To achieve this we use *textual segments* as 'date', 'location', 'epitaph' which describe the arrangement of the inscription. The knowledge of the *textual segments* allows for the determination of *reference systems, references* and *meanings*. *Reference systems* and *references,* as opposed to textual or editorial annotations, constitute a conceptual framework for corpus annotation which are important for cross-cultural and cross-linguistic comparisons.

XML (Bray et al. 2004) is, without question, the best supported annotation meta-language. To our knowledge, however, no XML language for tombstone corpora has been developed so far. EpiDoc for example, aims at the annotation of Epigraphs (Anderson et al 2007) and developed a rich scheme for the textual elements on the basis of TEI (Sperberg-McQueen & Burnard 2002).

However, EpiDOC yet does not provide an annotation framework beyond the text, such as the description of graves, graveyards or their ethnic and religious environments. In addition, EpiDoc stresses the individuality of the object, given the function of the epitaphs as revelation of the individual personality (cf. Edgette 1989/1992).

A corpus, however, serves a different purpose. In a corpus the individual stone, as well as any other individual feature is meaningless. When annotating a corpus, we annotate only those features which, beyond the purposes of data management and data retrieval, enter a system of meaningful oppositions (cf. Fages 1968). In terms of statistics, a feature is not annotated as long as there is no conjecture of a correlation with another feature. TSML is thus basically designed , for the annotation of features of correlation, developed on the basis of our experience with the massive annotation of Taiwanese tombstones and some tombstones from other countries.

2.1.TSML, Basic Structure

Simplicity, uniformity and flexibility of TSML is achieved by using the element in combination with a type-attribute as shown in Figure 2. We do not specify any constraints on the hierarchy of div-types as there are tombs without graveyards, tombs within tombs, graveyards without tombs, tombs in a church, tombstones without tombs and a tombstone-side, for example as photo, without the stone. Symbols, images, photos, maps can be at all levels, as well as texts which can be within the images, on the grave or on the tombstone.

```
tsml>
<div type='graveyard' north='55.34254' east='13.55456'religion='christianism'>
  <media mime_type='image/jpg' src='http://....../>
  <div type='church' background_color='green' set_up='1954-10-05' />
  <div type='graveyard_section' ethnicity='ami people'>
   <div type='tomb' direction='180' background_color='red' orientation='downhill'
        set_up='1962-09-02'>
     <div type='tomb_side' side='inside' vertical='90' direction='270>
       <div type='image' description='fish'/>
     </div>
     <div type='tombstone'>
      <div type='version' set_up='1962-09-02' status='lost'/>
      <div type='version' set_up='2002-01-05'>
        <media mime_type='image/jpg' src='http://....../>
        <div type='tombstone_side' background_color='white' foreground_color='black'
             writing_direction='t2br2l' script='han-zi'>
         <div type='photo' floating='top' size='6cm' description='male'>
           <media mime_type='image/jpg' src='http://....../>
         </div>
         <div type='symbol' floating='top' size='10'
              description='presbitarian cross'/>
         <div type='text' floating='top'>
          <div type='text' floating='right'>
           <div type='p'>text goes here</div>
          </div>
          <div type='text' floating='right' language='ami' script='katakana'>
           <div type='p'>text goes here</div>
```

Figure 2: The basic XML-structure of TSML based on ***div***-elements and ***type***-attributes, describing here an imaginary tomb. *Type*-attributes cascade from *div* to *div*.

Table 1: Values of the type-attribute of the div-element in TSML.

	Explanation
graveyard	Site where tombs are located.
graveyard_section	Graveyard sections may relate to different ethnicities or religions.
church, temple, ...	A building related to cults which contains a grave or is located in a graveyard.
tomb	A site containing the remains of one or more deceased.
tomb_side	An inner or outer wall of a tomb.
tombstone	The tombstone as 3-dimensional object.
tombstone_side	A 2-dimensional view on the tombstone.
tombstone_unit	Relatively independent units within a tombstone or a tombstone-side.
text, p, w, c, stroke	Containing mainly text.

image	Containing mainly an image
symbol	Containing mainly a non-figurative symbol
photo	Containing mainly a photo

For all div-elements, attributes are assumed to be inherited (to cascade) from the mother div-element to the daughter div-element in the absence of a more specific value. The information contained by some of these attributes cannot be seen on the tomb or tombstone directly and must be inferred or measured from other sources (GPS, compass, map, archives).

Table 2: Attributes of the div-element to be inherited from mother-div to daughter. Attributes marked '*' have been suggested in this or similar way in Debartolo Carmack 2002.

	Examples	Explanation
name*	Taipei Fude	Graveyards, graves may have official or unofficial names.
description		Free text input
location*	Taipei	Name of town, city, township where the entity is located.
caretaker*	Taipeishi	
caretaker address*		Caretakers might be contacted for additional information.
composition*	marble	Basic material: *marble, slate, granite, sandstone, limestone, metal, brick, concrete, ceramics*.
status*	abandoned	Useful to explain data loss, data endangerment. Values: *abandoned, maintained, overgrown, eroded, broken, lost*.
north	5,88789	Latitude as decimal WGS84 datum (cf. NIMA 97).
east	52,87465	Longitude as decimal WGS84 datum (cf. NIMA 97).
elevation	417	The elevation above mean see level in meter.
direction	90	Cardinal direction: *0=360*=North, *90*=East, *180*=South, ...
orientation	downhill	Non-compass directional system: *uphill, downhill, upcoast, downcoast, upstre am, downstream, landward, seaward, lakeward, mountainward, streetward, concentric*.
side	inside	*Inside, outside* with respect to the outer border of an object.
vertical	90	*90*=vertical, *0*=horizontal, wall and roof respectively.
set-up	2001-09-01	Time of construction/ building/ writing/ photographing.
floating	right	Relative position within the mother div-element, observer position opposite to the orientation, as in CSS absolute position (Bos et al. 2007). Alternative values: *right, left, top, bottom*.
display	block	Display according to CSS (Bos et al. 2007). Alternative values: *block, inline, list-item, superimposed, none.*
background- color	red	The color of the background, as in Çelik and Lilley 2003.
foreground-color	red, green, ...	The color of the foreground, as in Çelik and Lilley 2003.
religion	Buddhism, ...	The main religious orientation according to XNLRDF[1].
ethnicity	Hakka, Ami	The main ethnicity according to XNLRDF.
language	eng, deu	ISO 639-3 language codes, cf. XNLRDF.
writing-direction	t2bl2r, l2rt2b t2br2l, ..	Top-to-bottom left-to-right (Chinese), right-to-left top-to-bottom (English), top-to-bottom left-to-right (Mongolian), cf. XNLRDF.
script	Latin, Arabic	The set of characters or signs used according to XNLRDF.
nb_of_tombs	301	The number of tombs in this div.
nb_of_tombstones	417	The number of tombstones in this div.

2.2.Balanced Data

A corpus should be balanced (Biber et al. 1998). Although for a tombstone corpus, criteria of balancedness might be better to define than for a text corpus, e.g. collecting one photo per 1000 tombs, balancedness through sampling is impossible to achieve. Nobody knows all graveyards and how many tombs there are and those that we find may be inaccessible or decayed. In addition, naïve balancedness is not what we want. We want a tombstone corpus to have different

[1] XNLRDF, the Natural Language Research Description Framework, cf. Streiter et Stuflesser 2006.

granularities for different subset of the data. Graveyards of minorities are photographed and annotated exhaustively. In relatively uniform Han-communities of major cities samples are taken. Under uniform sampling conditions, no comparison inside smaller groups would be possible. We therefore introduce weights to achieve a numerical balancedness. A weight, stipulates how many items of the population are represented by one sampled item. Using the given, estimated or interpolated values or $inhabitants_{ly}$ and $life\text{-}expectancy_{ly}$ for a locality *l* and a year *y*, we estimate $population\text{-}of\text{-}tombs_{ly} \approx inhabitants_{ly}/life\text{-}expectancy_{ly}$. Then, $population\text{-}of\text{-}tombs_{ly}/sample\text{-}size_{ly}$ *yields* the weight for the graves of a location. The weight of these graveyards for larger geographic or administrative units can be derived by multiplying the weight of the smaller unit with the quotient of $population\text{-}of\text{-}tombs_{larger\ unit}/population\text{-}of\ \text{-}tombs_{smaller\ unit}$. Additional refinements in this calculus handle different locations sharing one graveyard and, graveyards associated with locations of different hierarchical levels. To sum up, although the details of this model can be refined this kind of **calculated balancedness based on census data** with the possibility to have different granularities in different sub-corpora is conceptually superior to the sampling for text corpora.

```
<div type=graveyard name='anping old street, both sides of the road'>
 <weights>
  <weight  year='2005'  loc='anping'  value='289'/>
  <weight  year='2005'  loc='tainan city'  value='3468'/>
  <weight  year='2005'  loc='tainan city + tainan county'  value='6936'/>
  <weight  year='2005'  loc='taiwan'  value='55488'>
 </weights>
```

Figure 3: The weights assigned in TSML to each tombstone of a graveyard: One sampled tombstone of the year 2005 in Anping represents 289 assumed tombstones in Anping of the same year and 3468 tombstones in Tainan City of the same year.

2.3. Lost Data

Like any historic document, tombstones or images may become unreadable. If something is totally unreadable, this has to be marked as unreadable as opposed to not yet annotated. If something is partially readable, say like the given name *Deb?rah,* where *?* stands for the unreadable, we almost for certain recognize the name, but we would falsify the data if we would write that we have `read` *Deborah.* It would equally be suboptimal to encode the entire name as unreadable, as this data loss cannot be made up by every corpus user, e.g. for a foreign language like Hebrew: דְּבוֹרָה. In TSML we thus keep track of what the interpretations of the decaying traces are. Interpretations can be complemented with probabilities.

As readability, especially of Chinese characters or Egyptians hieroglyphs might vary below the level of a character at the level of a radical or stroke, we cannot rely on character indices as for the annotation of the American National Corpus (cf. Ide & Romary 2006). If, for example, we can read a vertical stroke in a position where we expect a Chinese number, we might interpret this as 一(1) or if an additional scratch we see is a vertical stroke, as 十(10). Such phenomena can be described by the use of the <analysis> and <interpretation> element in combination.

```
v type='s'>He eats fish
nalysis type='3ps'>
div type='w' include='yes'>He</div>
div type='c' include='no'> </div>
div type='verb_root' include='no'>eat</div>
div type='infl_morph' include='yes'>s</div>
div type='c'> </div>
div type='w'>fish</div>
analysis>
iv>
```

Figure 4 (left): An example of how an discontinuous structure is annotated with the help of the *analysis*-element. **Figure 5 (right)**: An example of how different interpretations are derived from different analyses.

The <analysis> element provides a syntagmatic analysis of the mother element while repeating its content. Those elements marked as *include='yes' '*belong to the type analyzed (*<analysis type='xxx'*, here the markers of a 3rd person singular subject). The <interpretation> element list possible paradigmatic choices. Preferences in their selection are marked by *selected='yes'/'no'* or *probability='0.8'*.

```
<div type='p'>min guo 5
        <interpretation>
                <div type='c' language='' writing-direction='' probability='' selected='yes'>1
                        <analysis type='c'>
                                <div type='stroke' include='yes' >--</div>
                        </analysis>
                </div>
                <div type='c'    language='' writing_direction='' probability=''
selected='no'>10
                        <analysis type='c'>
                                <div type='stroke' include='yes'    display='superimpose'>-
-</div>
                                <div type='stroke' include='yes'
display='superimpose'>|</div>
                        </analysis>
                </div>
        </interpretation>
</div>
```

Fig 2: This tombstone became unreadable in it's lower right part. Which characters can be seen and which can be guessed?

At this level we describe single characters, as shown here for the hypothetical case that the first character of a date might be read as 1 or 7 and the second is completely unreadable.

2.4.Meaningful Data

The content of the div-elements may have a *reference.* The name of a person and a photo may have the same reference. Sometimes tombstones are bilingual, containing in two languages the same references.

The *references,* like times, persons and places are entities which across time lead an imagined or real existence independently from the grave or the tombstone and which are referred or alluded to in the tomb or tombstone. Relations among references are treated as references. Temporal references are obtained by translating the date we find in a specific calendar uniformly into the corresponding date of a calendar of reference. This calendar of reference might or might not be among the reference systems. In the same way we can map the names of a city onto a reference system of imagined or real cities, or personal names to a reference repository of persons. The interest in annotating references derives from the historical, geographical or sociological facts they reflect. Such facts might be apprehended through tombstones or they provide background information for the interpretation of other data on the tombstones. The references to time on the tombstone, for example, can be used to analyze historical developments in the form and content of tombstones.

We identify *reference systems* with social mediators which shape experience and awareness for members of a culture. For a description of mediation as psychological process see, among others, Wertsch 1988. Thus, for some researchers looking at tombstones, neither the exact date (which we will call *reference)*, nor the exact wording (which we will call *meaning)* of a date might be

interesting, but instead what kind of calendar is used. The calendar is a social construct which mediates psychological processes. Similarly, researchers might be interested in whether symbols taken from a Christian symbol repository or from a Jewish symbol repository instead of the symbol used (the *meaning*) or what the symbol means (the *reference).* The way that people are referred to, or the language and the writing system, all are additional, all too obvious reference systems that merit an analysis. We hypothesize that whichever reference system is used, it reflects the social and psychological reality of the community and the different reference systems (calendar, language, names etc) do not cooccur randomly.

In our research on Taiwanese tombstones, still another reference system, that of the local origin is of central importance. Many Taiwanese families actually have the possibility to chose between the T*anghao,* a mythological place name in North China, a place name in South China from where the ancestors immigrated (Jiguan) or the place name in Taiwan where the family lived (Taiwan diming). The name might not be important for an analysis, the reference system of the origin however hints on identities communities maintain. In addition, we expect the reference system (tanghao/jiguan/taiwan diming) to cluster with other reference systems, such as the calendar (Japanese/Chinese/Republican/Gregorian).

http://140.127.211.213/img/tombo/xishu2007-06-13/dsc00785.jpg
http://140.127.211.213/img/tombo/3--2007-09-24/dsc07455.jpg
http://140.127.211.213/img/tombo/3--2007-09-24/dsc07457.jpg

Figure 3: Left to right. Three local reference systems, the Tanghao, the Jiguan (place name in China), the place name in Taiwan.

http://140.127.211.213/img/tombo/12--2007-07-11/dsc02679.jpg
http://140.127.211.213/img/tombo/13--2007-07-11/dsc02660.jpg
http://140.127.211.213/img/tombo/9--2007-07-09/dsc02162.jpg
http://140.127.211.213/img/tombo/10--2007-07-12/dsc03237.jpg

Figure 2: Left to right, top to bottom. Four calendars found on Taiwanese tombstones and their references: Japanese calendar, the traditional Chinese calendar, the Republican calendar and the Gregorian calendar.

Another, important category for analysis are *meanings*. *Meanings* derive from *reference types*, where *reference types* are intentional abstractions of references, for example, 'person', 'father', 'date', 'date of birth' etc as they can be apprehended from the DTD. Thus, while 'Bill, father of John' has a reference which represents this fact, 'X, father of Y' has a reference type, here that element in the DTD that describes 'father of'-relations. Reference types are annotated if we want to analyze and compare the different meaningful components of a tombstone (symbols, words, expressions, arrangement) across regions, cultures and languages. If we annotate these meaningful component with a reference type, for example, 'father of', we can access all meaningful expressions with the reference type in their context. Note, that this model requires all quasi-equivalent expressions, which in corpus-linguistic approaches might be simply defined by a synset, to be defined in relation to *references* within the XML This might be too strong a claim and imply that that one has to invoke the apparatus of the reference system during annotation, even if the reference system allows for one reference only. We can tweak this by making the *reference type* obligatory and the *reference system* optional as in *** ***, in ** ** or in *** ***

The relation between the *reference type, reference* and the *meaning* can be described as follows. The *meaning* of a component without a reference derives from its *reference type* ('honorific') and the difference between this component (*** ***) and all others with the same *reference type.* If there is a reference, the

meaning of a component derives from its *reference* ('9') and the difference between this component (*** ***) and all others with the same *reference.*

In addition to linguistic expressions, symbols, colors or arrangements may have references and thus can be grouped into reference types. During the annotation process the color of a segment *** *** might be elaborated into *** .*** Given the formulaic nature of tombstones, meanings and references are not necessarily transparent, even for members of that linguistic or cultural community, and thus have to be annotated.

3. Conclusion

Just more text.

And more text in second indented paragraph.

4. References

Biber, D., Conrad, S., Reppen, R. 1998. *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge University Press.

Bos, B., Çelik, T., Hickson, I. & Lie, H.W. 2007. Cascading Style Sheets Level 2 Revision 1 (CSS 2.1) Specification. W3C Candidate Recommendation 19 July 2007. URL: http://www.w3.org/TR/CSS21/cover.html, accessed 24.9.07.

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler E., Cowan, J., Yergeau, F. eds., 2004. *Extensible Markup Language (XML) 1.1*, W3C. URL: http://www.w3.org/TR/xml11, accessed 22.6.07.

CES, Corpus Encoding Standard, Part 5: Encoding Linguistic Annotations. 2000. URL: http://www.cs.vassar.edu/CES/CES1-5.html, accessed 22.6.07.

Çelik, T. & Lilley, C. 2003. CSS3 Color Module, W3C Candidate Recommendation 14 May 2003. URL: http://www.w3.org/TR/css3-color, accessed 24.9.07.

Clark, E.W. 1989/1992. The Bigham Carvers of the Carolina Piedmont: Stone Images of an Emerging Sense of American Identity. In: R.E. Meyer ed., *Cemeteries & Gravemarkers, Voices of American* Culture. Utah State University Press.

Debartolo Carmack, S. 2002. *Your Guide to Cementary Research*, Betterway Books.

Edgette, J.J. 1989/1992. The Epitaph as Personality Revelation. In: R.E. Meyer ed., *Cemeteries & Gravemarkers, Voices of American* Culture. Utah State University Press.

Elliott, T. ed., 2007. EpiDoc: Guidelines for Structured Markup of Epigraphic Texts in TEI. URL: http://www.stoa.org/epidoc/gl/5/, accessed 22.6.07.

Fages, J.B. 1968. *Comprendre le structuralisme*. Privat, Toulouse.

Frege, G. 1892. Sinn und Bedeutung. In: *Zeitschrift für Philosophie und philosophische Kritik*, NF 100, 1892, S. 25-50.

Frembgeni, J.W. 1989. Religious Folk Art as an Expression of Identity: Muslim Tombstones in the Gangar Mountains of Pakistan. In: *Muqarnas*, Vol. 15, pp. 200-210.

Good, J. & Hendryx-Parker, C. 2006. Modeling Contested Categorization in Linguistic Databases. In: *2006 EMELD Workshop on Digital Language Documentation, Tools and Standards, the State of the Art*. Michigan State University in East Lansing, Michigan, June 20-22.

Ide, N., Romary, L. 2006. Representing Linguistic Corpora and Their Annotations. Proceedings of the *Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy.

The Hindu. Buddhist Tombstone discovered at Chittayam, *The Hindu, online edition of India's National Newspaper* 8.7.2005, URL: http://www.hindu.com/2005/07/08/stories/2005070811950300.htm, accessed 23.6.07.

NIMA Technical Report TR8350.2, "Department of Defense World Geodetic System 1984, Its Definition and Relationships With Local Geodetic Systems", Third Edition, 4 July 1997. URL: http://earth-info.nga.mil/GandG/publications/tr8350.2/tr8350_2.html, accessed 24.9.07.

Rath, G. 1986. Hebrew Tombstone Inscriptions and Dates. In: *Chronicles (Newsletter of the Jewish*

Genealogical Society of Philadelphia), Vol. 5, No. 1 (Spring 1986), pages 1-4.
de Saussure, F. 1916/1995. *Cours de linguistique générale*, éd. Payot.
Sperberg-McQueen, C.M. & Burnard, L. eds., 2002 *Guidelines for Text Encoding and Interchange.* University of Oxford.
Statistical yearbook of the Republic of China, 2002.
Streiter, O. & Stuflesser, M. 2006. Design Features for the Collection and Distribution of Basic NLP-Resources for the World's Writing Systems. In: *Intl. Workshop Towards a Research Infrastructure for Language Resources*, LREC Workshop, Genova, Italy, 22 May 2006.
Wertsch, J.V. 1988. *Vygotsky and the Social Formation of Mind*. Harvard University Press.
Wurm, S. ed., 2001. *Atlas of the World's Languages in Danger of Disappearing*. Paris, UNESCO.